Source versus spectral cues in the perception of indexical features in speech

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for Graduation with Distinction in Speech and Hearing Science in the Undergraduate Colleges of The Ohio State University

By: Hannah E. Ortega

The Ohio State University

April 2015

Project Advisors: Dr. Robert A. Fox and Dr. Ewa Jacewicz, Department of Speech and Hearing Science

ACKNOWLEDGMENTS

I would like to thank Dr. Robert Fox and Dr. Ewa Jacewicz for all their help and wisdom, not only with this thesis project, but throughout the past couple of years. As a team they have guided me through this project and I have learned so much. I am very thankful to have had them as my mentors.

I would also like to thank Jessica Hart and Makenzie Laase for making the arduous process of writing a thesis fun! You gals are non-stop laughs and I am so grateful to have known you both!

TABLE OF CONTENTS Page

LIST OF TABLES AND FIGURES

TABLES

FIGURES

ABSTRACT

Spoken language includes two different forms of information: linguistic (message related) and indexical (related to individual speaker characteristics). Recent research revealed that listeners are sensitive to indexical features such as regional dialect spoken in their speech community. However, little is known how listeners form a perceptual representation of speaker identity and how the indexical information is conveyed by the vocal source (related to voice) and filter (related to the changing shape of the vocal tract during speech production). This study explores the nature of the acoustic cues that listeners may use to identify gender and dialect of speaker. Nine spontaneous utterances produced by 40 speakers (20 male, 20 female) from two different regional dialects spoken in central Ohio (OH) and western North Carolina (NC) were the stimuli. These utterances were equally divided into three sets. One set was unprocessed (except for amplitude equalization). A second set was low-pass filtered at 400 Hz, retaining voice and prosodic information, but little content. A third set was processed through an 8-channel noise vocoder eliminating all harmonic information related to the vocal source (analogous to cochlear implant processing). These three stimulus sets were played to 20 OH listeners who indicated whether the token was produced by a man or a woman, from OH or NC. Gender identification rates were high (means > 89%) across all three conditions with clear > LP filtered > vocoded. The rates for dialect identification were significantly lower overall with the LP-filtered condition close to chance (58%). Discussion centers on listener use of acoustic cues and perceptual sensitivity (d') to gender and dialect. This research provides preliminary data for future assessment of sensitivity to these indexical properties by listeners with cochlear implants.

Chapter 1

INTRODUCTION

Speech represents a complex acoustic pattern. Shaped by the coordinated movements and actions of the articulators (e.g., tongue, jaw, lips) and the pulmonic system (lungs and vocal folds), spoken language encodes two different forms of information that is used by the listener: linguistic information related to the message of the signal (e.g., phonemes, words, syntax, semantics) and indexical information about the speaker (Levi & Pisoni, 2007; Clopper & Bradlow, 2009). Abercrombie (1967) divided up these indexical properties into three basic sets of information about the speaker: (1) group memberships (e.g., regional, dialectal and social aspects), (2) physical characteristics (e.g., age, gender, size) and (3) mental states (e.g., fatigue, amusement, anger, suspicion, etc.). Interest in the indexical features in the speech literature has been growing in the past ten years.

Recent research has shown that listeners are sensitive to indexical features such as regional dialect spoken in their speech community (Clopper et al., 2006; Jacewicz & Fox, 2012). However, little is known how listeners form a perceptual representation of speaker identity and how the indexical information is conveyed by the vocal source (related to voice) and filter (related to the changing shape of the vocal tract during speech production). This study looks more carefully at the acoustic cues that are utilized by a listener when identifying the gender and dialect of a speaker, two indexical properties of speech.

Gender differences are cued primarily by voice characteristics (Skuk & Schweinberger, 2014). The perceived pitch of male voices is lower than the pitch of female voices (average fundamental frequency (F0) is about 100-120 Hz for adult males and 200-220 Hz for adult

females). Listeners are able to use this pitch (voice) information to help determine gender of a speaker.

The voice also transmits rich emotional and habitual cues about speaker (e.g., intonation, affect, anger, speech tempo, rhythm, pauses). These indexical cues are responsible for significant variations in F0. It is unknown how much information about regional dialect is contained in a speaker's voice. This study modifies acoustic information available to the listener to examine how voice information may give a listener cues about dialect of a speaker.

Dialect features are primarily conveyed by acoustic spectral cues such as vowel formant pattern, stop consonant releases, consonant cluster reduction or occurrence of r-colored vowels (Clopper & Pisoni, 2004). However, little is known how much information about speaker gender is contained in spectral cues alone (related to the vocal tract filtering function). This study further investigates this question by removing much of the voice information, retaining the spectral information, and presenting stimuli in the vocoded condition. The results will give more insight into how listeners use spectral information when identifying speaker gender.

Although this study will use normal hearing (NH) participants, this research will also serve to provide preliminary data for future assessment of sensitivity to these indexical properties by listeners with cochlear implants (CI). For that reason, vocoded speech will be used which provides input to the NH listener that mimics the electroacoustic stimulation that CI users receive while listening to speech.

Chapter 2

METHODOLOGY

Participants

Twenty-one participants (14 female, 7 male) between 21 and 30 years of age were subjects in this study. One participant's data were discarded due to a low correct response rate in the clear speech condition. The mean age of the remaining participants was 23.35 years (SD= 2.33). Participants were recruited in by word of mouth. Fifteen participants were full-time graduate or undergraduate students at The Ohio State University; the others were employed full-time outside the University. All participants had lived in Central Ohio (here defined as the geographic areas within a one hour drive from Columbus) for the majority of their lives and spoke the dialect of American English common to this region (Midland). Two of the listeners had undergone speech therapy for one year as young children. Only three of the participants had lived outside of Ohio for more than one year and the average time spent living in Ohio was 19.7 years for all twenty participants. All reported normal hearing. Subjects performed the task in November and December of 2014.

Procedure

Participants were asked to identify the gender and dialect of 40 different speakers (20 from Central Ohio and 20 from western North Carolina). Each heard utterances from these 40 different speakers over Sennheiser 640 headphones in a sound attenuating booth. After hearing an utterance the participants indicated if they thought the speaker of the utterance was from Central Ohio or western North Carolina, male or female. They made their selection on a computer by clicking (using a mouse) on one of four response boxes displayed on the computer monitor in front of them.

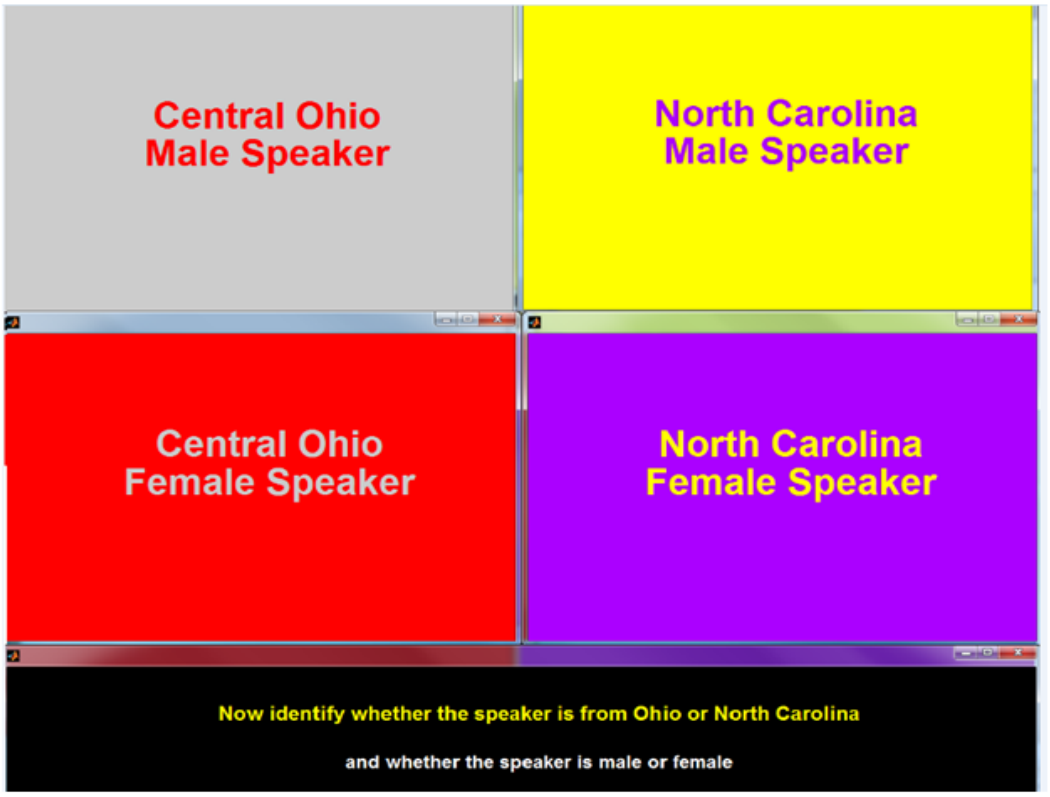

Figure 1. A screen shot of response boxes used by the participant during the experiment to indicate the geographic region and gender of the speaker.

Each participant was asked to fill out a background form that contained questions about his/her speech, language, dialectal, and educational background (See Appendix). This experiment was conducted under a protocol approved by the Institutional Review Board of the Office of Research at Ohio State.

The experimenter verbally explained to the participants as they were sitting in the sound-attenuating booth that they would be listening to many utterances spoken by different talkers from Ohio and North Carolina, both male and female. They were asked to make a decision about the gender and dialect of the speaker of the utterance they heard and to indicate that decision by clicking one of four response boxes that appeared on a computer screen. The experimenter then

left the booth. Participants heard stimuli under each of three stimulus conditions (clear speech (CS), low pass-filtered (LP), and vocoded speech (VS) which will be described in the following section. In each condition, they were presented with ten practice utterances followed by 120 test utterances. The practice utterances were only to ensure that the listener understood the task and that the utterances were being presented at a comfortable volume for the listener. Participants were allowed to ask questions or express any concerns following each of the practice trials. After each utterance was presented, the participants chose one of the four response boxes (Central Ohio Male, Central Ohio Female, North Carolina Male, and North Carolina Female) to indicate the dialect and gender of the speaker of the utterances they just heard. If the participant was unsure of the answer, he/she was instructed to make their best guess. The order of presentation of stimulus conditions was counterbalanced across the study. In particular, half of the participants heard the utterance conditions in the order: CS, LP, VS, while the other half heard the condition in the order CS, VS, LP.

The test in its entirety was completed in 45-60 minutes and participants were compensated $15 for their time. Compensation came from a scholarship received by the College of Social and Behavioral Science that supports student research projects.

Stimuli

The speech stimuli were selected from a database of previously recorded spontaneous conversations of 40 speakers, 20 from Central Ohio (OH) and 20 from western North Carolina (NC). Details of the original recordings can be found in Jacewicz, Fox & Wei (2010). There were 10 male speakers and 10 female speakers in each dialect group, Speakers ranged in age from 52-68 years. For OH, the mean ages were $M = 57.7$ ($SD = 3.4$) for males and $M = 60.8$ (SD

= 5.9) for females; for NC, *M* = 58.8 (*SD* = 5.9) for males and *M* = 59.4 (*SD* = 2.8) for females. Nine utterances from each speaker were selected to serve as test utterances, for a total of 360 utterances.

In choosing the utterances to be used in the experiment, the conversational speech sample from each separate speaker was reviewed by an experimenter using acoustic analysis software (Adobe Audition). The experimenter then extracted nine utterances from each speaker based on the amount of syllables within the utterance and its overall intelligibility. The experimenter was also careful to ensure the utterances selected fell within the conversational pauses of the speaker so as not to interrupt the natural speech pattern of that speaker. The selected samples did not contain any lexical information that might strongly link a speaker to any region (e.g., "I like huntin' an fishin'). Additionally, male/ female- specific information was not included. For example, sentences such as: "My wife went to the store," was not included as this might imply this utterance was spoken by a male. The chosen utterances were defined as shorter (<= 8 syllables) or longer (>8 syllables) and were relatively equally divided across the experimental conditions described below.

Table 1. The number of shorter and longer utterances in each of the three conditions (CS, LP, VS).

Condition	**Number of Shorter Utterances**	**Number of Longer Utterances**
Clear Speech	66	54
Low-pass filtered	62	58
Vocoded Speech	66	54

Table 2. Number and percentage of total utterances and those defined as "shorter" and "longer".

Total Utterances (#, %)	Shorter Utterances (#, %)	Longer Utterances (#, %)
360, 100 %	194, 53.9 %	166, 46.1%

The three different experimental conditions included clear speech, low-pass filtered, and vocoded speech. In the clear speech condition the stimuli represented the original unprocessed waveforms (except for amplitude equalization). In the low-pass filtered condition all stimulus sentences were low-pass filtered so that all spectral information above 400 Hz was removed—these low-pass filtered versions were then amplitude equalized to match the level of the clear speech tokens. Low-pass filtering retains voice (prosodic) information, but little or no spectral detail (and no semantic/syntactic content). In the vocoded condition stimulus utterance were processed through an 8-channel noise vocoder. The summary of each channel's parameters (start, end and width) is provided in Table 3. Noise-vocoded speech eliminates voice (harmonic) information but retains much of the spectral envelope detail that provides semantic/syntactic content. The intelligibility of vocoded speech increases with the number of channels, and 8 channels are considered sufficient for good speech intelligibility (Louizou et al., 1999). Vocoded speech simulates cochlear implant processing (Friesen et al., 2001).

Table 3. Start, end and width of each of the eight channels (in Hz) used in creating the vocoded speech.

Channel	Start	End	Width
1	300	477	177
2	477	722	245
3	722	1061	339
4	1061	1528	467
5	1528	2174	646
6	2174	3066	892
7	3066	4298	1232
8	4298	8000	3702

Noise-vocoded speech eliminates voice (harmonic) information but retains much of the spectral envelope detail that provides semantic/syntactic content. The intelligibility of vocoded speech increases with the number of channels, and 8 channels are considered sufficient for good speech intelligibility (Louizou et al., 1999). Vocoded speech simulates cochlear implant processing (Friesen et al., 2001).

The stimuli were placed into each of these 3 conditions randomly. There were 3 utterances from each speaker in each condition, for a total of 120 stimuli in each condition. No individual utterance appeared in more than a single experimental condition.

Figures 2-4 display a spectrogram for the utterance “There’s a lots of shopping opportunities” (this utterance was not among the test items).

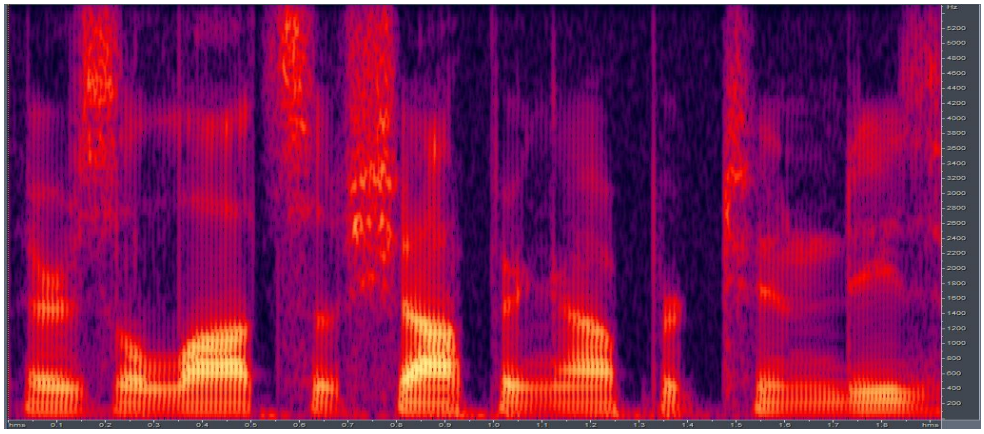

Figure 2. A spectrogram of a clear speech stimulus utterance. It is acoustically unprocessed except for amplitude equalization.

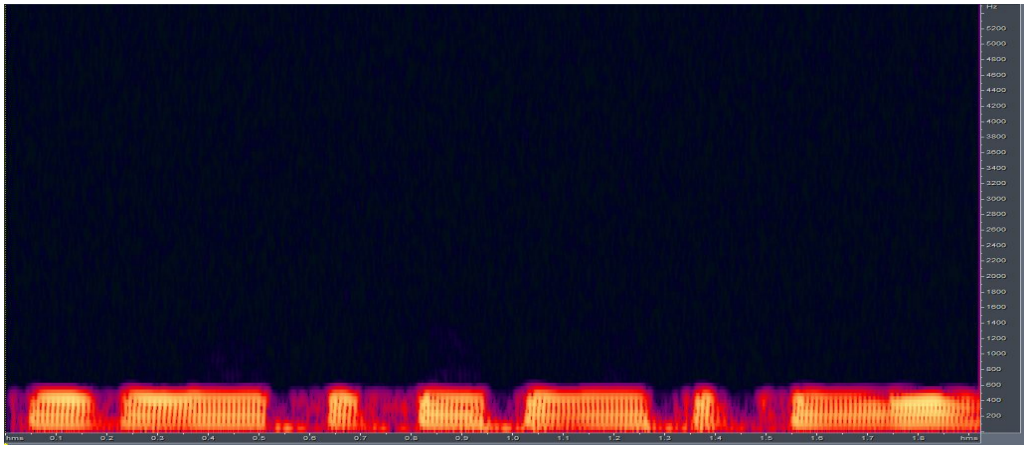

Figure 3. A spectrogram of a low-pass filtered utterance in which all spectral information above 400 Hz was removed (and the altered token amplitude equalized).

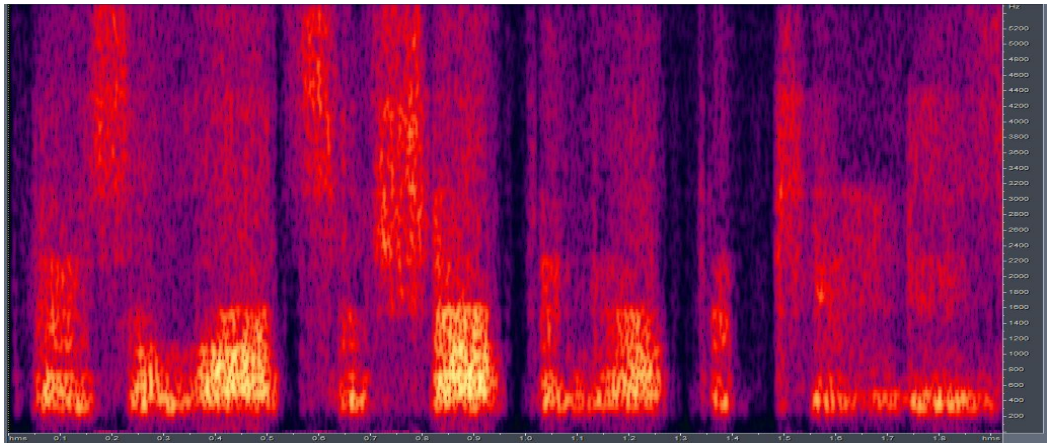

Figure 4. A spectrogram of a vocoded stimulus processed through an 8-channel noise vocoder (and amplitude normalized). Note that the spectral detail in the form of "formant patterns" can be seen to resemble those found in the clear-speech version.

Chapter 3

RESULTS

Participants' responses were analyzed using signal detection theory (STD). As is well known, accuracy alone is not a good indicator of performance because it does not separate two components that contribute to accuracy: sensitivity and bias. In addition, accuracy cannot easily account for a listener's decisions made under different degrees of stimulus uncertainty (Lynn & Barrett, 2014). STD allows for separating listeners' sensitivity (d') to dialect from their response bias (C) (Green & Swets, 1966; Lynn & Barrett, 2014; Macmillan & Creelman, 2005). In this analysis, the correct categorization of an OH talker was "hit" and the categorization of a NC talker as an OH talker was a "false alarm." The sensitivity measure, d', was calculated for each individual listener. Listener sensitivity to both gender and dialect recognition were analyzed. The results are described below and displayed in Figures 5-8.

Gender recognition

A two-way analysis of variance (ANOVA) was used to assess sensitivity (d') to speaker gender (i.e., the ability to detect whether the speaker was a male or a female) with the within-subject factors stimulus condition (CS, LP, VS) and speaker dialect (OH, NC). The analysis showed that the main effect of listening condition was significant [$F(2, 38) = 113.3$, $p < .001$] and that listeners were most sensitive to gender in CS, followed by LP and VS condition, respectively. Subsequent t-tests showed that the differences between all three conditions were significant. In the CS condition, with both voice and spectral information, listeners were able to recognize the gender of the speaker with nearly perfect accuracy. In the LP condition, with mainly voice information, listeners performed better at gender recognition than in the VS condition with mainly spectral information present. Overall, in all conditions listeners were relatively sensitive to gender of the speaker.

The main effect of dialect was not significant [$F(1,19) = 2.37$, $p = .141$]. However, there was a significant interaction between condition and dialect [$F(2, 38) = 7.1$, $p = .006$]. In CS, listeners were equally sensitive to gender when hearing OH or NC speakers. In LP, sensitivity was greater for OH speakers than for NC speakers. However, in VS, sensitivity was greater for NC and not for OH speakers. The results of the gender recognition test are displayed in Figures 5 and 6.

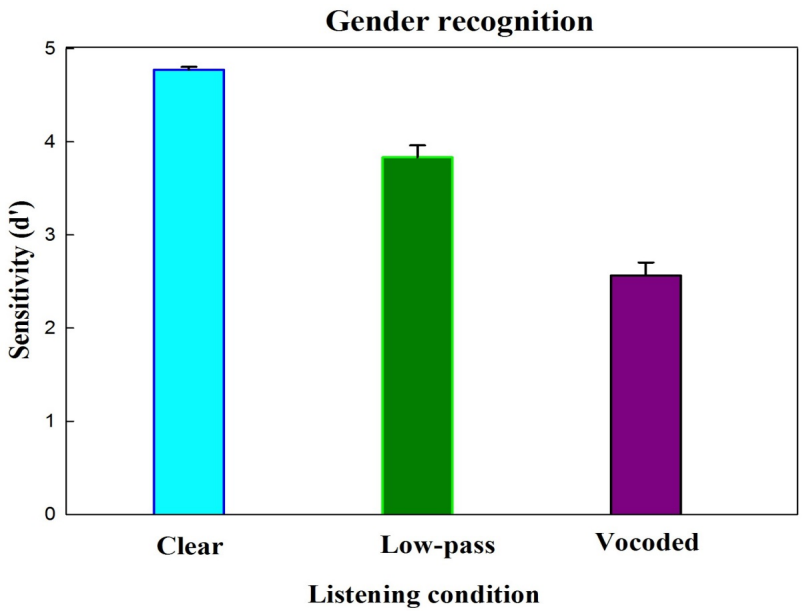

Figure 5. Mean *d'* values for gender decisions for the CS, LP and VS conditions.

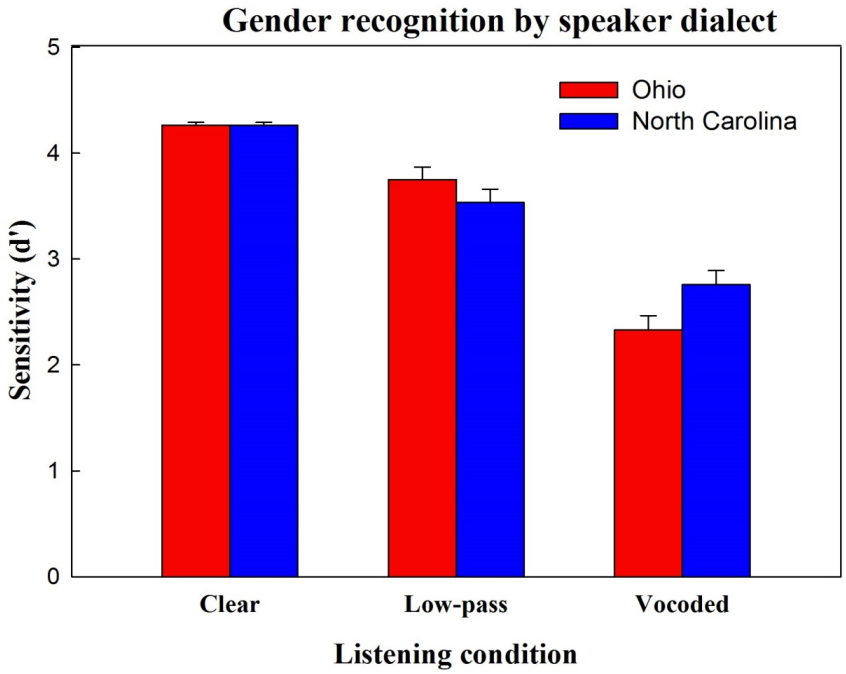

Figure 6. Mean *d'* values for gender decisions for the three experimental conditions broken down by speaker region.

Dialect recognition

Next the *d′* values based on dialect decisions were analyzed using a two-way ANOVA with the within-subject factors condition and speaker gender. The main effect of condition was significant [$F(2, 38) = 105.6$, $p < .001$]. Participants were most sensitive to dialect differences in CS, followed by the VS and LP conditions, respectively. Subsequent *t*-tests showed that the differences among all three conditions were significant. There was also a significant main of effect of gender [$F(1, 38) = 8.9$, $p = .008$], which showed greater dialect sensitivity in response to male ($M = 1.43$) than female ($M = 1.21$) speakers. A significant gender by condition interaction arose as listeners were more sensitive to dialect when listening to male speakers in CS and LP but in VS, they were more sensitive when listening to female speakers [$F(2, 38) = 35.9$, $p < .001$]. The results of the dialect recognition test are displayed in Figures 7 and 8.

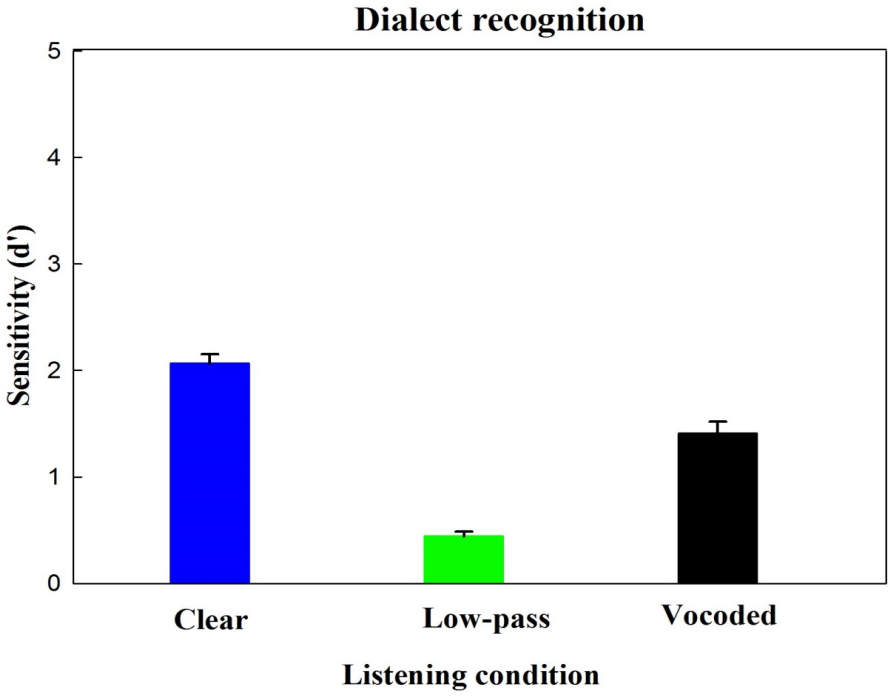

Figure 7. Mean *d′* values for dialect decisions for the CS, LP and VS conditions.

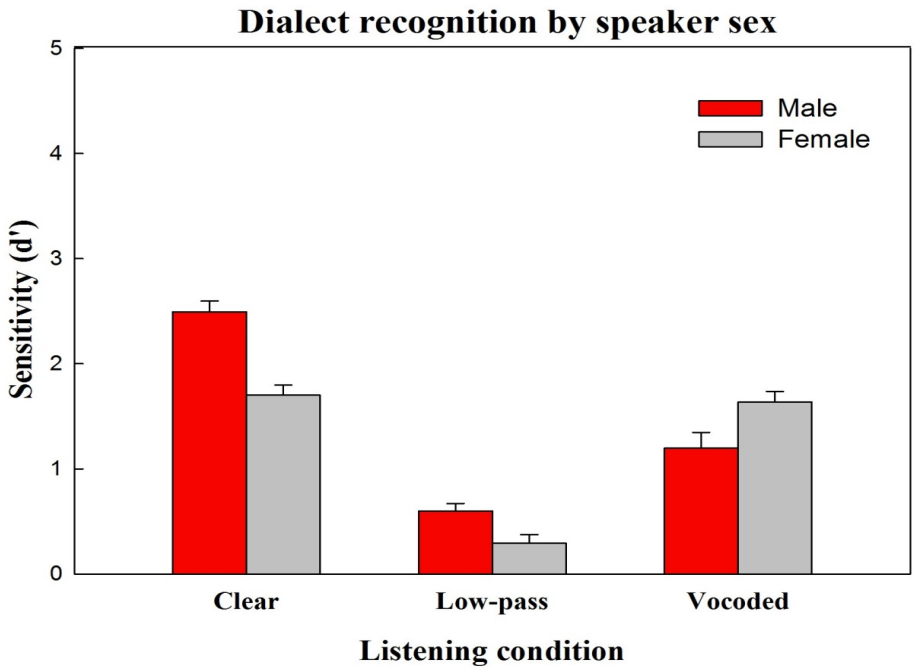

Figure 8. Mean *d′* values for dialect decisions for the three experimental conditions broken down by speaker gender.

Chapter 4

DISCUSSION

The analysis of gender recognition showed that overall listeners can recognize the gender of a speaker with relative ease, whereas dialect recognition is a more challenging task. Because sensitivity was relatively high in all conditions regarding gender recognition, gender information is conveyed not only by the speaker's voice (LP condition), but also when voice cues are unavailable (VS condition). Further, when only voice cues are present (LP condition), listeners recognize gender better when the speaker has the same dialect as the listener. That is, all of our subjects who spoke with a central Ohio dialect better recognized gender when listening to a speaker who also spoke with a central Ohio dialect. However, when voice cues are absent (VS condition) listeners are more sensitive to speakers with a non-native dialect.

The analysis of dialect recognition showed that listeners are best at this task when all speech cues are present, as in the CS condition. Because sensitivity to dialect was the lowest for listeners in the LP condition, it appears that voice information (LP condition) contributes little to dialect identification. It appears that spectral cues (VS condition) give the listener the most information about dialectal features. Further, when voice cues are available (CS and LP conditions), listeners are more sensitive to dialect in response to male speakers. However, when voice cues are absent (VS condition) female speech gives the listener more information about the speaker's dialect.

As mentioned, vocoded speech simulates cochlear implant (CI) processing as determined by the match in performance between normal hearing listeners presented with 8-channel vocoded speech and successful CI listeners (Friesen et al., 2001). The listener performance in the vocoded speech condition provides information about the extent to which CI users might have access to the same acoustic information and thus, might be able to process these same indexical cues. We hope this information can be applied clinically to help professionals who work with CI users to better understand what speech information CI users are getting from their device and thus, what modifications to speech/language therapy might need to be considered.

BIBLIOGRAPHY

Abercrombie, D. (1967). *Elements of General Phonetics.* Chicago: Aldine.

Clopper, C.G. & Bradlow, A.R. (2009). Free classification of American English dialects by native and non-native listeners. *Journal of Phonetics*, *37*, 436-451.

Clopper, C., & Pisoni, D. (2004). Some acoustic cues for the perceptual categorization of American English dialects. *Journal of Phonetics, 32*, 111-140.

Clopper, C., Levi, S., & Pisoni, D. (2006). Perceptual similarity of regional dialects of American English. *Journal of the Acoustical Society of America, 119*, 566-574.

Friesen, L. M., Shannon, R. V., Baskent, D., et al. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *Journal of the Acoustical Society of America*, 110, 1150-1163.

Jacewicz, E. and Fox, R. A. (2012). The effects of cross-generational and cross-dialectal variation on vowel identification and classification. *Journal of the Acoustical Society of America* 131, 1413-1433.

Levi, S.V. & Pisoni, D.B. (2007). Indexical and linguistic channels in speech perception: Some effects of voiceovers on advertising outcomes. *Psycholinguistic Phenomena in Marketing Communications* (T. M. Lowrey, ed.). Lawrence Erlbaum Associates. pp. 203-219.

Loizou, P.C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *Journal of the Acoustical Society of America, 106*, 12097-2103

Skuk, V.G., & Schweinberger, S.R. (2014). Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. *Journal of Speech, Language, and Hearing Research, 57*, 285-296.